An NCME Instructional Module on Using Differential Step Functioning to Refine the Analysis of DIF in Polytomous Items

Randall D. Penfield and Karina Gattamorta, *University of Miami*, and Ruth A. Childs, *Ontario Institute for Studies in Education of the University of Toronto*

Traditional methods for examining differential item functioning (DIF) in polytomously scored test items yield a single item-level index of DIF and thus provide no information concerning which score levels are implicated in the DIF effect. To address this limitation of DIF methodology, the framework of differential step functioning (DSF) has recently been proposed, whereby measurement invariance is examined within each step underlying the polytomous response variable. The examination of DSF can provide valuable information concerning the nature of the DIF effect (i.e., is the DIF an item-level effect or an effect isolated to specific score levels), the location of the DIF effect (i.e., precisely which score levels are manifesting the DIF effect), and the potential causes of a DIF effect (i.e., what properties of the item stem or task are potentially biasing). This article presents a didactic overview of the DSF framework and provides specific guidance and recommendations on how DSF can be used to enhance the examination of DIF in polytomous items. An example with real testing data is presented to illustrate the comprehensive information provided by a DSF analysis.

Keywords: differential item functioning, polytomous items, graded response model

Randall D. Penfield is an Associate Professor, School of Education, University of Miami, PO Box 248065, Coral Gables, FL 33124-2040; penfield@miami.edu. His primary research interests include educational measurement, differential item functioning, and item response theory. Karina Gattamorta is a doctoral student in the Research, Measurement, and Evaluation Program, School of Education, University of Miami; kgattamorta@miami.edu. She is interested in educational measurement and student assessment. Ruth A. Childs is an Associate Professor, Department of Human Development and Applied Psychology, Ontario Institute for Studies in Education, University of Toronto; rchilds@oise.utoronto.ca. Her research interests include item response theory, test preparation and administration, and response processes.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Information regarding the development of new ITEMS modules should be addressed to: Dr. Mark Gierl, Canada Research Chair in Educational Measurement and Director, Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, University of Alberta, Edmonton, Alberta, CANADA T6G 2G5.

Let us consider a test being developed to assess proficiency in a particular content domain for individuals having a range of racial, ethnic, economic, and social backgrounds. At some point in the process of developing the test and validating the obtained scores, the question of fairness must be addressed. The test developer must provide a justifiable argument that the decisions and interpretations made as a result of the test scores lead to outcomes that are fair and equitable across examinees of varying racial, ethnic, economic, and social backgrounds. In the technical measurement literature (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), fairness has been defined according to four properties: a fair test is one for which (a) the items are free of bias, (b) all examinees are provided equal opportunity to demonstrate their level of proficiency in the intended construct, (c) all examinees have had an equal opportunity to learn the content being assessed (with the exception of employment, credentialing, and admissions testing), and (d) the distributions of test scores are as equal as possible across different groups of examinees. Of the four properties of fairness, that of item bias has garnered the greatest attention in

the technical measurement literature (Camilli, 2006). Contemporary methods for studying item bias are based, in part, on the framework of differential item functioning (DIF). An item contains DIF if individuals having the same level of proficiency, but belonging to different groups, have a different expected response to the item (Holland & Thayer, 1988; Penfield & Camilli, 2007; Roussos & Stout, 2004). The presence of DIF, along with content-based evidence concerning the causes of the DIF, provides evidence that the item may be biased (Camilli, 2006).

Incipient procedures for assessing DIF focused on dichotomous items (Camilli & Shepard, 1994; Holland & Wainer, 1993; Penfield & Camilli, 2007; Roussos & Stout, 2004). However, the increasing use of polytomous item formats (e.g., performances that can be scored according to the degrees of proficiency, rather than as simply correct or incorrect) has led to the development of numerous methods for assessing DIF in polytomous items (Penfield & Camilli, 2007; Penfield & Lam, 2000; Potenza & Dorans, 1995). Tests of DIF in polytomous items address whether individuals having the same level of proficiency, but belonging to different groups, have the same chance of obtaining each score level of the polytomous response variable. A limitation of traditional measures of DIF for polytomous items is that they provide only an item-level index of the DIF effect (or an item-level test of the null hypothesis of no DIF) and thus provide no information concerning which score levels are implicated in the DIF effect or whether some score levels are implicated more than others. For this reason, traditional DIF measures for polytomous items can be conceptualized as omnibus measures of DIF. Because omnibus measures of DIF provide no information concerning which score levels are manifesting the DIF effect, they provide limited information to help guide the identification of specific components of the item manifesting the DIF effect and the potential causes of the DIF effect.

The limitations of omnibus DIF measures make clear the need for a DIF methodology that examines measurement equivalence in relation to each score level of the polytomous item. The probability of observing each score level of a polytomous item is defined according to a series of step functions describing the chance that an individual will progress, or step, from one score level to a higher score level (e.g., the step from a score of 1 to a score exceeding 1, the step from a score of 2 to a score exceeding 2, etc.). It is the properties (i.e., underlying parameters) of these step functions that ultimately dictate the probability of observing each score level for an individual with a particular level of ability (Baker, 1992). As a result, an examination of the between-group difference in measurement properties in relation to each score level can be pursued through an examination of the between-group difference in the properties of the step functions underlying the polytomous item. This framework has been referred to as differential step functioning (DSF; Penfield, 2006, 2007). The framework of DSF provides a mechanism for examining the between-group difference in measurement properties at each step, thus providing detailed information concerning where along the polytomous response process a lack of measurement equivalence may exist for the groups under consideration.

The framework of DSF provides DIF analysts with several advantages over the omnibus measures of DIF. First, tests of measurement invariance based on the DSF effects can be more powerful than the omnibus DIF tests when the magnitude and/or sign of the DSF effect varies across the steps of the underlying polytomous response variable (Penfield, 2006, 2007). In the extreme case where the sign of the DSF effect changes across the steps (i.e., is positive for one step but negative for another), the power of DSF-based tests of invariance has been shown to be more than 10 times that of the omnibus tests of DIF (i.e., a power of .045 for the omnibus test of DIF compared with a power of .85 for the test of DSF; Penfield, 2006). A second advantage of the DSF framework is that it allows the DIF analyst to pinpoint precisely which score levels (or steps) are responsible for an observed DIF effect. That is, if a polytomous item is flagged for DIF, then the analysis of DSF can be used to isolate the components of the item that require further content review and possible revision and ultimately suggest the factors causing the DIF. Because the identification of the causes of DIF is the key to decisions about item revision and/or removal (Bolt, 2000; Douglas, Roussos, & Stout, 1996; Gierl & Khaliq, 2001; Oshima, Raju, Flowers, & Slinde, 1998; Scheuneman, 1987; Schmitt, Holland, & Dorans, 1993; Swanson, Clauser, Case, Nungester, & Featherman, 2002), the framework of DSF can play a pivotal role in such decisions. In addition, the growing interest in the consideration of cognitive strategies used in responding to items (DiBello, Roussos, & Stout, 2007; Leighton & Gierl, 2007; Mislevy, 2006) places a new emphasis on understanding between-group differences in measurement properties in relation to these strategies. DSF provides a mechanism for identifying between-group differences in strategies underlying the responses to polytomous items.

To date, the only accounts of DSF and related methodology have been technical and have provided limited guidance on the use and interpretation of DSF results. In this article, we present a nontechnical overview of the DSF framework and available methodology for assessing DSF and provide recommendations for the use and interpretation of DSF analyses. Issues of particular importance include: (a) how the results of a DSF analysis can help target investigations into the causes of DIF, (b) what methods can be used to evaluate DSF, (c) what criteria should be used to flag large DSF magnitudes, and (d) how DSF analyses can be most effectively used in conjunction with traditional DIF analyses. In addition, we illustrate the use of DSF using a real data set.

What Is DSF?

Let us begin the description of DSF by considering a polytomously scored item having four score levels with outcomes denoted by Y. In this example, the score levels will be assigned the values of 0, 1, 2, and 3. Polytomous item response theory (IRT) models can be used to specify the probability that an examinee with a particular level of ability will obtain each of the possible score levels (0, 1, 2, and 3, in this case) for the item in question. These IRT models, however, are simply a reformulation of a series of relatively simple models, called *step functions*, which describe the probability that an examinee with a particular level of ability steps (or advances) from one score level to a higher level. For a polytomous item having r score levels, there will be $J = r - 1$ step functions. Thus, in our example of a polytomous item having four response options ($r = 4$), there are three underlying step functions ($J = 3$).

There are several different forms of step functions, and the different forms correspond to distinguishing and defining characteristics of polytomous IRT models. For example, the graded response model (GRM; Samejima, 1969, 1972, 1997) uses the cumulative form, and the partial credit family of models (Masters, 1982; Muraki, 1992, 1997) uses the adjacent-categories form. The distinguishing properties of different forms of step functions are described in numerous sources (Agresti, 1990; French & Miller, 1996; Penfield & Camilli, 2007). In this article, we focus on the cumulative step function form, whereby the jth step function describes the probability that an examinee successfully advances to a score equal to or greater than j (it is the "equal to or greater than" aspect to which the cumulative label is attributable). In our example of a four-category polytomous item, the cumulative form would lead to the following step functions: (a) the first step function describes the probability that an examinee advances from a score of 0 to a score of at least 1, that is, the probability that $Y \geq 1$; (b) the second step function describes the probability that an examinee advances from a score of 1 or lower to a score of at least 2, that is, the probability that $Y \geq 2$; and (c) the third step function describes the probability that an examinee advances from a score of 2 or lower to a score of 3, that is, the probability that $Y = 3$.

Each of the J step functions is defined using a two-parameter logistic model such that each model contains a difficulty parameter (b_j) that is specific to the step and a discrimination parameter (a) that is constant across all steps. Recalling that the jth step function specifies the probability of observing a response greater than or equal to j [i.e., $P(Y \geq j)$], as a function of ability (θ), the jth step function has the form

$$P(Y \geq j \mid \theta) = \frac{\exp[a(\theta - b_j)]}{1 + \exp[a(\theta - b_j)]}. \quad (1)$$

Note that the parametric form of the jth step function is equivalent to the two-parameter logistic IRT model (Lord, 1980) commonly applied to dichotomously scored (e.g., multiple-choice) items. The parameter a dictates the discriminating power of each step function (i.e., how strongly each step discriminates between individuals of low vs. high levels of ability). The parameter b_j dictates the difficulty of the jth step, and the specific value of b_j represents the value of ability required to have a probability of .5 (i.e., a 50% chance) of successfully advancing at the jth step. Under the GRM, it is assumed that the values of b_j are ordered from lowest to highest across the successive steps.

An illustration of the three cumulative step functions underlying a four-category polytomous item is shown in Figure 1. These step functions have the parameter values $a = 1, b_1 = -1.5, b_2 = -.5, b_3 = 2.0$. The first step function is located furthest to the left, indicating that for any level of ability the probability of successful advancement at the first step is always higher than that for the second and third steps. Similarly, the third step function is located furthest to the right, indicating that for any level of ability the probability of successful advancement at the third step is always lower than that of the first and second steps. It is also relevant to note that the relative spacing between the three step functions is not identical—although the second step function is not very far to the right of the first step function, the third step function lies relatively far to the right of the second step function. This indicates that the difficulty of a successful transition at the second step is not substantially more than the difficulty at the first step, but the difficulty of a successful transition at the third step is considerably more than the difficulty at the second step. The only requirement placed on the relative locations of the step functions is that each successive step function be located to the right of the previous step function (i.e., each successive step increases in difficulty).

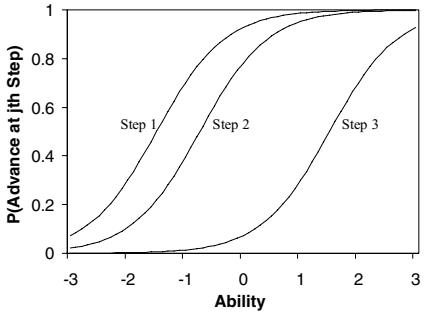

FIGURE 1. Trace lines for the three step functions underlying a hypothetical four-category polytomous item.

The framework of DSF is concerned with the extent to which the properties of each step function differ between the reference (R) and focal (F) groups, and thus the extent to which the parameters $b_1, b_2, \ldots, b_J$ differ between the groups. Because the a parameter is not specific to the step (i.e., it is constant across all steps), most of the emphasis in examining DSF falls on the between-group differences in the b_j parameters (the rationale for this is addressed in a later section). To describe this more formally, let us denote the step function difficulty parameters of the two groups by b_{jR} and b_{jF}, respectively. Then, the identification of DSF concerns the extent to which $b_{jR} = b_{jF}$ for each step. The condition of no DSF for the jth step holds if $b_{jR} = b_{jF}$.

Figure 2 presents a pictorial representation of DSF for which b_{jF} is .3 units higher than b_{jR} for each of the three steps of the polytomous item (i.e., $b_{jF} = b_{jR} + .3$), causing each of the step functions of the focal group to be shifted to the right of those of the reference group by .3 units. As a result, reference and focal group members who share the same level of ability do not share the same probability of successfully advancing at each step; the probability of

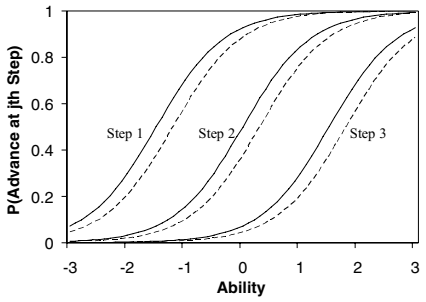

FIGURE 2. Trace lines for the three step functions underlying a hypothetical four-category polytomous item for the reference group (solid lines) and focal group (dashed lines). All steps display a constant DSF effect.

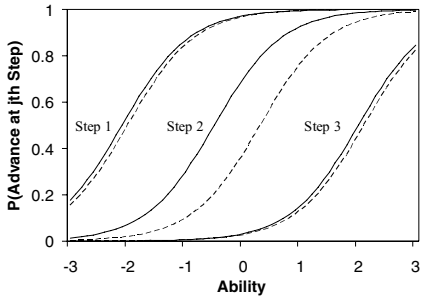

FIGURE 3. Trace lines for the three step functions underlying a hypothetical four-category polytomous item for the reference group (solid lines) and focal group (dashed lines). Step 2 displays a large DSF effect, and steps 1 and 2 display a negligible DSF effect.

advancing is higher for the reference group than for the focal group. This situation represents a relative advantage for the reference group at each step. The DSF presented in Figure 2 is characterized by a magnitude that is equal across all steps (the DSF effect is .3 for each step), and thus represents a *constant* form of the DSF effect.

The constant DSF form depicted in Figure 2 is but one of the many different DSF forms that are possible. The DSF effects can vary across the steps. For example, the DSF effect may be zero for all steps but one, or the DSF effect can be negative for one step but positive for another step. Figure 3 displays the situation in which the DSF effect varies across the three steps of a four-category polytomous item. In particular, the first and third steps have negligible DSF effects (for both these steps, $b_{jF} = b_{jR} + .1$), and the second step displays a large DSF effect ($b_{3F} = b_{3R} + .8$). An interesting property of the DSF framework is that the pattern of DSF effects can provide valuable information concerning the location and causes of the DIF effect. A description of particularly revealing patterns of DSF, and how the patterns can guide investigations into the causes of the DIF effect, is provided in a subsequent section.

The astute reader who is well versed in DIF methodology for dichotomous items will no doubt have noticed the similarity of DSF to the concept of DIF for dichotomously scored items. In the case of the dichotomous item, DIF concerns whether individuals having the same level of ability, but belonging to different groups, have the same probability of correct response. In contrast, DSF for the jth step concerns whether individuals having the same level of ability, but belonging to different groups, have the same probability of successfully advancing at the jth step. As such, each step can be heuristically conceptualized as a dichotomous item, and the presence of DSF is examined at each of the J steps using a framework that is analogous to the one examining DIF in dichotomous items. It should come as no surprise, then, that methods used for examining DSF are similar to those originally developed for use with dichotomous items (these methods are described in a subsequent section).

Having introduced the concept of DSF, we are now able to describe the link between DSF and DIF. Because the properties of the polytomous item as a whole are determined by the J step functions, we can conceptualize the DIF effect as the aggregated DSF effect across the J steps. If the condition of no DSF holds for each step of a polytomous item, then it must be the case that the condition of no DIF holds for that item. Similarly, as the DSF effects increase in magnitude, we expect the DIF effect to increase. Despite the direct link between DSF and DIF, there are several circumstances in which many DIF indices are expected to be relatively insensitive to large DSF effects. One such circumstance is when a large DSF effect exists in one, and only one, step. In this case, the large DSF effect will be diluted by the zero DSF effects of the other steps, yielding a relatively small aggregated DIF effect. A second circumstance is when two large DSF effects have opposite signs (one step displays DSF favoring one group, and a different step displays DSF favoring the other group), thus yielding a net DIF effect that is near zero. In these cases, the framework of DSF provides a more accurate representation of measurement invariance than the omnibus measures of DIF.

Making Score-Level Interpretations from Step-Level Results

Upon observing a substantial DSF effect for a particular step, the DIF analyst is faced with the task of identifying which specific score levels are responsible for the DSF effect. That is, step-level DSF effects must be translated into score-level properties. For example, a large DSF effect for the second step of a polytomous item containing four score levels indicates that the transition from the lowest two score levels to the highest two score levels (i.e., assuming a cumulative step function form) was relatively more difficult for one group. Because the relative advantage was associated with the transition to the highest two score levels, the DSF effect alone does not provide direct information concerning which of the highest two score levels are responsible for the DSF effect. It may be that only the third score level is responsible, or it may be that both the third and fourth levels are responsible.

Although translating the DSF effects into meaningful score level interpretations is not completely transparent, several strategies can be used to make informed judgments as to the score levels containing a potentially biasing factor. First, if a substantial DSF effect exists in one isolated step, say the jth step, then it is likely the case that the factor responsible for the DIF effect resides in the jth score level. This is because the DSF effect at the jth step indicates that a between-group difference exists in advancing to any score greater than or equal to the jth score level, suggesting that the between-group difference is attributable to a property of the jth score level. Similarly, if substantial DSF effects exist in several adjacent steps, say steps j and $j + 1$, then it is likely that the factor responsible for the DIF effect resides in score levels j and $j + 1$.

Second, if the DSF effects are relatively constant in magnitude across all steps, then there exists evidence that the factor responsible for the DSF effect is an item-level property that may be located in the content of the item stem (e.g., a writing prompt, or the stimulus for a performance task) or the general properties of the performance task itself. Because this effect is expected to impact all steps, we can conceptualize this as an item-level effect. The following section describes in more detail how examination of the pattern of DSF effects can be used to make

judgments concerning the location of potential causes of the DIF effect.

Using the Pattern of DSF Effects to Help Identify the Cause of DIF

As described in the previous sections, the presence of a DSF effect in a particular step can help the DIF analyst in targeting the specific score levels manifesting a potentially biasing factor. We can, however, make even more use of the DSF effects in understanding the causes of DIF through an analysis of the pattern of the DSF effects across the J steps of the polytomous item. In particular, the specific pattern of the DSF effects across the J steps of the polytomous item can help guide the analyst in identifying the possible cause(s) of the DIF effect and in making a decision about item revision or removal.

Although there are an infinite number of patterns that the J DSF effects can assume, several general groupings of patterns are particularly revealing of the causes of the DIF effect. Penfield, Alvarez, and Lee (2009) described these groupings within a two-dimensional taxonomy of DSF patterns. The first of these dimensions distinguishes between pervasive and nonpervasive DSF. Pervasive DSF is observed when all J steps display a substantial DSF effect, and thus the DSF effect is *pervasive* across all score levels. The presence of pervasive DSF suggests to the analyst that the cause of DIF is exerting its influence at the item level. For example, pervasive DSF may be observed in a writing task where students are asked to respond to a particular prompt. In such an item, the presence of pervasive DSF would imply that the factor responsible for the lack of invariance is inherent in the content of the prompt itself. In contrast, nonpervasive DSF exists when only one or a few steps display a substantial DSF effect. The presence of nonpervasive DSF implies that the factor causing DIF may be isolated to one or a few steps. For example, consider a writing task in which DSF appears only in a score level that requires well-structured paragraphs, in addition to the characteristics required by the scoring criteria for the lower score levels. In this case, the nonpervasive DSF provides evidence that the DIF effect is not necessarily due to content in the writing prompt but rather is isolated to properties of the particular level pertaining to paragraph structure. Making this distinction between pervasive and nonpervasive DSF can prove valuable in determining whether the cause of DIF is due to an item-level property or a property of one or more particular score levels.

The second dimension of the DSF taxonomy pertains to the consistency of the DSF effects across impacted steps, distinguishing between constant, convergent, and divergent forms of DSF. Constant DSF is observed when the steps displaying a DSF effect are relatively equal in magnitude and sign. Although constant pervasive DSF provides evidence that the factor responsible for the DSF effect is a property of the item, constant nonpervasive DSF indicates the factor responsible for the DSF is restricted to the affected score levels and thus is not necessarily an item-level property. Convergent DSF describes the situation in which affected steps display a DSF effect of the same sign (i.e., favoring the same group) but different magnitude, providing evidence that the causal factors are manifested differentially across steps. It may be the case that an item-level effect impacts score levels differently, or more than one biasing factor is present. Divergent DSF is characterized by affected steps displaying opposite signs, meaning that the relative advantage shifts between groups across the steps. The presence of divergent DSF implies that the causes of the DSF effects are different for the affected score levels, and thus more than one causal property is at play. Identifying the presence of divergent DSF is of paramount importance because many DIF statistics are expected to be relatively insensitive when divergent DSF effects cancel one another at the item level, yielding a net DIF effect near zero.

Statistical Methods for Evaluating DSF

As is the case for the investigation of DIF, several different statistical approaches can be applied to the investigation of DSF. To date, three general approaches for evaluating DSF have been described in the literature: an IRT approach, an odds ratio approach, and a logistic regression approach. In general, these approaches follow some of the same methodology employed in assessing DIF in dichotomous items, with the exception that: (a) the polytomous response variable for the studied item must first be dichotomized to create the associated J step-level response variables, and (b) a separate analysis must be conducted for each of the J steps. As discussed previously, there are several methods that can be used to create the J step-level response variables, and we focus here on the cumulative approach because this has been the most widely adopted in the literature.

The IRT approach for investigating DSF is based on examining the between-group differences in the b_j parameters of the jth step function (described by Equation 1) for the studied item. It follows that the DSF effect for the jth step can be defined as the between-group difference in the b_j parameter, given by:

$$\Delta(b_j) = b_{jF} - b_{jR}. \tag{2}$$

The situation of $\Delta(b_j) = 0$ corresponds to no DSF for the jth step, the situation of $\Delta(b_j) > 0$ corresponds to a relative advantage for the reference group on the jth step, and the situation of $\Delta(b_j) < 0$ corresponds to a relative advantage for the focal group on the jth step. The example of DSF presented in Figure 2 corresponds to $\Delta(b_1) = .3$, $\Delta(b_2) = .3$, and $\Delta(b_3) = .3$. The example of DSF presented in Figure 3 corresponds to $\Delta(b_1) = .1$, $\Delta(b_2) = .8$, and $\Delta(b_3) = .1$.

In practice, the values of b_j for the reference and focal groups must be estimated, and thus an estimate of the DSF effect for the jth step is given by the difference between the estimated values of b_j for the reference and focal groups. It has been shown that $\Delta(b_j)$ is equal to the signed area between the jth step functions for the two groups (Cohen, Kim, & Baker, 1993), which is equivalent to Raju's area measure of DIF applied to dichotomously scored items (Raju, 1988). As a result, the magnitude of the DSF effect for the jth step may be interpreted on a similar metric as the DIF effects defined using Raju's area measure for dichotomous items.

Tests of the null hypothesis of no DSF under the IRT approach can be conducted using two different approaches. The first approach is to divide $\Delta(b_j)$ by its estimated standard error to form a test statistic that is distributed

approximately as standard normal. This approach is equivalent to Lord's (1980) one degree of freedom chi-square statistic used for assessing uniform DIF in dichotomous items, as applied to each step. The second approach for testing the null hypothesis that the DSF effect associated with the jth step equals zero uses a likelihood ratio test (Ankenmann, Witt, & Dunbar, 1999; Kim & Cohen, 1998; Thissen, Steinberg, & Wainer, 1993), whereby the likelihood obtained when the item parameters are constrained to be equal for both groups (the compact model) is compared with the likelihood obtained when the step parameters are freed to vary between the two groups (the augmented model). Descriptions of the likelihood ratio tests for the analysis of DIF using an IRT approach are numerous (Camilli & Shepard, 1994; Penfield & Camilli, 2007; Thissen et al., 1993). Although the IRT approach provides a comprehensive parametric framework for examining DSF, its appropriate implementation is dependent on having adequately large sample sizes within each step and data that adequately fit the respective step function. IRT approaches to examining DSF can be conducted using a variety of IRT calibration programs, including BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003), IRTLRDIF (Thissen, 2001), and MULTILOG 7 (Thissen, Chen, & Bock, 2003).

A second approach for examining DSF is to employ the odds ratio method described by Penfield (2006, 2007). This method evaluates the DSF effect of the jth step by comparing the odds of successfully advancing at the jth step for the reference and focal group members having the same level of ability (where ability is typically approximated using an observed score variable such as the total summated test score). The between-group equality of the odds of successful advancement at the jth step can be assessed by considering the ratio of the odds of success of the reference group over that of the focal group, yielding what is known as an odds ratio. The typical value of the odds ratio taken across all levels of ability is widely referred to as a common odds ratio, and the natural logarithm of this common odds ratio for the jth step is denoted by λ_j (the log-odds ratio possesses a metric that is symmetric about 0, which is preferable to the asymmetric odds ratio metric). A value of $\lambda_j = 0$ corresponds to no DSF at the jth step, a value of $\lambda_j > 0$ corresponds to DSF favoring the reference group for the jth step, and a value of $\lambda_j < 0$ corresponds to DSF favoring the focal group for the jth step. The common log-odds ratio for the jth step can be estimated using the common log-odds ratio DSF effect estimator described by Penfield (2006, 2007). The resulting estimator, $\hat{\lambda}_j$, is analogous to the Mantel–Haenszel common log-odds ratio estimator (Mantel & Haenszel, 1959) widely used in the assessment of DIF in dichotomous items (Holland & Thayer, 1988).

The odds ratio approach holds an interesting relationship to the DSF effect defined under the IRT approach. Assuming that the IRT step function defined in Equation 1 fits the data for the jth step, and that the observed test score is an adequate approximation for ability, then the following approximate relationship holds:

$$\lambda_j \approx a[\Delta(b_j)], \tag{3}$$

where the symbol $\approx$ denotes an approximate equivalence due to the difference in the ability variable used for the IRT and odds ratio approaches. That is, the common log-odds ratio is approximately proportional to the between-group difference in the b_j values for the jth step, where the proportionality is determined by the item-level discrimination parameter, a. Details concerning the derivation of this relationship are discussed in Penfield (2007), and the relationship is well documented for the odds ratio and IRT approaches to modeling DIF effects in dichotomous items (Penfield & Camilli, 2007). The point to take from this is that, whether the DSF effect for the jth step is modeled under the IRT or odds ratio approaches, both approaches are estimating a similar effect.

The null hypothesis of no DSF at the jth step using the odds ratio approach can be tested using the following test statistic:

$$z(\hat{\lambda}_j) = \frac{\hat{\lambda}_j}{SE(\hat{\lambda}_j)}. \tag{4}$$

The test statistic shown in Equation 4 is distributed approximately as standard normal (Hauck, 1979).

Conducting a DSF analysis using the odds ratio approach can be accomplished using several software programs. Version 4.0 of the program DIFAS (Penfield, 2005) can compute $\hat{\lambda}_j$ and report the results of $z(\hat{\lambda}_j)$ for tests of the null hypothesis of no DSF. Alternatively, the analyst can use widely available statistical software packages (e.g., SPSS, SAS, Stata, R) to compute $\hat{\lambda}_j$ and $z(\hat{\lambda}_j)$ by: (a) creating a set of J dichotomously coded variables that coincide with the J cumulative step functions, and (b) computing $\hat{\lambda}_j$ for each step by obtaining the Mantel–Haenszel common log-odds ratio estimator for each of the J dichotomously coded variables using the summated test score (or some other appropriate measure of ability) as the stratifying variable. The advantage of using DIFAS over general statistical software packages to conduct DSF analyses is that DIFAS automatically creates the dichotomously coded step-level variables within the DSF analysis, thus obviating the analyst's need to conduct the dichotomizations prior to running the DSF analyses.

A third approach for investigating DSF is that of logistic regression, first introduced in the context of DIF detection in polytomous items by French and Miller (1996). The logistic regression approach for examining DSF is based on a step function model that posits the probability of successfully advancing at the jth step as a function of an observed test score variable (X) and group membership (G) using the following logistic equation:

$$P(Y \geq j \mid X) = \frac{\exp(\beta_{j0} + \beta_{j1}X + \beta_{j2}G)}{1 + \exp(\beta_{j0} + \beta_{j1}X + \beta_{j2}G)}. \tag{5}$$

The grouping variable, G, is typically a dummy coded variable (i.e., $G = 1$ for the reference group, and $G = 0$ for the focal group), and the coefficient β_{j2} corresponds to the DSF effect at the jth step. The condition of $\beta_{j2} = 0$ corresponds to no DSF at the jth step, the condition of $\beta_{j2} > 0$ reflects DSF favoring the reference group, and the condition of $\beta_{j2} < 0$ reflects DSF favoring the focal group. Although it is possible to augment the logistic regression equation with an interaction term (the interaction of X and G) that corresponds to the nonuniform DSF effect (French & Miller, 1996; Swaminathan & Rogers, 1990), we restrict our discussion here to uniform DSF effects. A discussion of nonuniform DSF effects is presented in a later section.

A useful relationship exists between the DSF effects defined under the logistic regression and odds ratio approaches.

If the logistic regression model for the jth step function shown in Equation 5 is used to compute the log-odds ratio at a particular level of X, then the resulting log-odds ratio reduces to β_{j2}, the DSF effect parameter for the jth step. This suggests that there is an equivalence of the estimated value of β_{j2} and the common log-odds ratio, λ_j. The estimation methods used in estimating β_{j2} and λ_j differ (logistic regression uses an iterative maximum likelihood approach, whereas the odds ratio approach is noniterative), and thus, although the estimated values of β_{j2} and λ_j will not necessarily be exactly identical, they will typically be very close to one another in value, as will be their estimated standard errors. This equivalence has been demonstrated in previous applications of these approaches (Alvarez & Penfield, 2007), and it suggests that extremely similar results will be obtained using the logistic regression and odds ratio approaches. Furthermore, the equivalence between β_{j2} and λ_j suggests that β_{j2} will be proportional to $\Delta(b_j)$. This proportionality is given by:

$$\beta_{j2} \approx a[\Delta(b_j)], \tag{6}$$

where the symbol $\approx$ denotes an approximate equivalence due to the difference in the ability variable used for the IRT and logistic regression approaches.

Testing the null hypothesis of no DSF using the logistic regression framework is conducted using a likelihood ratio approach that is similar to that used for IRT approaches to DSF and follows identical steps to those used in the application of logistic regression methods to the evaluation of DIF in dichotomously scored items (Camilli & Shepard, 1994; Penfield & Camilli, 2007). The analysis of DSF using the logistic regression approach can be conducted using widely available statistical software. The analyst must create the J cumulatively coded dichotomized variables (one for each of the J steps) within each polytomous item under investigation. For each analysis (and a separate analysis must be conducted for each step of each polytomous item), the statistical software (e.g., SPSS, SAS, Stat, R) will provide an estimate of β_{j2}, and also display -2(likelihood) of the specific model being run, which can then be used in computing the appropriate likelihood ratio tests.

The Concept of Nonuniform DSF

Despite the similarity of DSF to the study of DIF in dichotomous items, there is an important difference when considering the between-group difference in the a parameter. In dichotomous items, a between-group difference in the a parameter is easily interpretable as a nonuniform DIF effect—the relative advantage of one group changes across the ability continuum (Camilli & Shepard, 1994; Penfield & Camilli, 2007). The concept of a nonuniform DSF effect, however, is complicated by the mathematical forms underlying polytomous item response models that define the step functions described by Equation 1. For the majority of the widely used polytomous response models (i.e., graded response model, partial credit model, generalized partial credit model, rating scale model), the value of the a parameter is restricted to be constant across all J steps, and thus a unique nonuniform DSF effect cannot exist within each step (for an elaboration of this, as it pertains to DSF, see Cohen, Kim, & Baker, 1993). As such, under the IRT definition of DSF given above, the concept of a nonuniform effect is more relevant to the entire item than it is to unique step-level processes. Because the framework of DSF is based upon examining between-group differences in the measurement properties that are unique to each step, the between-group difference in the a parameter can be considered to fall outside the realm of the DSF framework. For this reason, we will focus our discussion of DSF effects on the between-group differences in the step-level difficulty parameters.

Although the definition of DSF provided in this article concerns only uniform effects (i.e., between-group differences in the step-level difficulty parameters only), the logistic regression approach has the flexibility to examine step-level nonuniform effects by including a term for the interaction of group membership (G) and observed test score variable (X). This approach is analogous to that described for examining nonuniform DIF effects in dichotomously scored items (Swaminathan & Rogers, 1990) and has been demonstrated in the application of logistic regression to DSF by French and Miller (1996). Other contingency table (e.g., odds ratio) approaches to examining nonuniform DIF in dichotomous items (Mazor, Clauser, & Hambleton, 1994; Penfield, 2003) could also be extended to the study of nonuniform DSF effects. However, the utility of examining nonuniform DSF has not been addressed in the literature, and it remains unclear how fruitful the study of nonuniform DSF effects would be in practice.

Criteria for Interpreting DSF Effects

The ultimate utility of DSF analyses is increased if small, medium, and large DSF effect sizes are defined. In general, the same criteria commonly used for categorizing the size of DIF effects in dichotomous items can be adopted for the categorization of DSF effects. The most widely used approach for classifying DIF magnitude is the ETS classification scheme based on the common log-odds ratio estimator (Zieky, 1993). Applying the effect size component of this classification scheme to $\hat{\lambda}_j$ (the step-level log-odds ratio estimator) yields the following criteria for the magnitude of the DSF effects: $|\hat{\lambda}_j| < .43$ corresponds to a small effect, $.43 \leq |\hat{\lambda}_j| < .64$ corresponds to a medium effect, and $|\hat{\lambda}_j| \geq .64$ corresponds to a large effect. The criteria of .43 and .64 used in defining the three categories have intuitive appeal because they correspond to odds of successful transition for one group that is approximately 1.5 times and 2.0 times that of the other group, respectively. Typically, steps falling into the large category warrant additional review of content for potentially biasing factors. However, examination of the pattern of the DSF effects to inform the location of the potentially biasing factor requires the consideration of the category associated with each step. Details of using the pattern of DSF effect categories to inform the location of the DIF effect are provided in Penfield et al. (2009).

Approaches for categorizing the DIF effects using logistic regression have been based on the difference in R^2 (R_d^2) obtained between the model with the grouping term (i.e., β_{j2} in Equation 5) and the model without the grouping term (Jodoin & Gierl, 2001; Zumbo & Thomas, 1996). These initial approaches proposed the following point values of R_d^2 to reflect DIF effect sizes: $R_d^2 = .02, .13$, and $.26$ represent small, medium, and large effects, respectively. We propose adapting these point values of R_d^2 to generate the following ranges of R_d^2 for classifying the DSF effects: $R_d^2 < .10$ corresponds to

a small DSF effect, $.10 \leq R_d^2 < .20$ corresponds to a medium DSF effect, and $R_d^2 \geq .20$ corresponds to a large DSF effect. It is also useful to note that the equivalence between β_{j2} and λ_j suggests that their estimates are expected to be very similar, and thus the ETS criteria described above for $\hat{\lambda}_j$ can be applied equally to $\hat{\beta}_{j2}$.

Criteria for interpreting the magnitude of the DIF effect using IRT approaches have not yet been firmly established in the measurement literature. However, based on the relationship between $\Delta(b_j)$, β_{j2}, and λ_j (see Equations 3 and 6), a sensible criteria can be established. As a general rule of thumb, we recommend the following criteria for DSF effects under an IRT model: $|\Delta(b_j)| < .25$ is a small effect, $.25 \leq |\Delta(b_j)| < .50$ is a medium effect, and $|\Delta(b_j)| \geq .50$ is a large effect. These criteria correspond roughly to those presented for $\hat{\lambda}_j$.

Using DSF Results to Test for DIF

Under the condition of no DSF for each of the J steps, it must be the case that no DIF exists. As a consequence, it is possible to evaluate the presence of DIF through the examination of the DSF effects. In particular, a test of the null hypothesis of no DIF can be undertaken through the test of the null hypothesis of no DSF at each of the J steps, whereby an adjusted Type I error rate (e.g., a Bonferroni-adjusted Type I error rate) is employed for each of the J tests of DSF to control for an inflated Type I error rate resulting from the multiple tests. If the null hypothesis of no DSF is retained for all J steps, then the null hypothesis of no DIF may be retained. If, however, the null hypothesis of no DSF is rejected for one or more steps, then the null hypothesis of no DIF can be rejected. This approach is referred to as the simultaneous step-level (SSL) test of DIF (Penfield, 2007) and was originally proposed in the context of the odds ratio DSF estimator. Because the SSL test is based on the step-level analyses, it does not require additional computation over and above those conducted for the analysis of DSF. In addition, the SSL test has been shown to be more powerful than other widely used tests of DIF (Penfield, 2007) when the magnitude and/or sign of the DSF effect varies across the J steps.

To gain a deeper understanding of the properties of the SSL test, it is necessary to distinguish between net and global tests of DIF for polytomous items. Net tests of DIF are based on the aggregation of the signed DSF effects across all steps, and thus can be insensitive to divergent DSF effects. Examples of net tests of DIF include Mantel's chi-square (Mantel, 1963), the Liu–Agresti cumulative common log-odds ratio estimator and associated test statistic (Liu & Agresti, 1996; Penfield & Algina, 2003), the standardized mean difference index and associated test statistic (Dorans & Schmitt, 1993), and the polytomous SIBTEST procedure (Chang, Mazzeo, & Roussos, 1996). In contrast, global tests of DIF consider the unsigned DSF effects across all steps, and thus are expected to be relatively sensitive under the condition of divergent DSF. The SSL test is a global test of DIF because it does not aggregate the signed DSF effects across the steps. As such, the SSL test is expected to be more sensitive than net tests of DIF when the DSF effects are divergent or vary greatly in their magnitude. Other global tests of DIF exist, such as the generalized Mantel–Haenszel chi-square statistic (Somes, 1986), and the relative performance of the SSL test to other approaches has not yet been explored in the literature. In general, net measures of DIF have been shown to be more sensitive than global measures when the DSF effects are constant across the steps, but global measures have been shown to be more sensitive than net measures when the DSF effects vary in sign and/or magnitude across the steps (Penfield, 2007; Wang & Su, 2004).

Using DSF and DIF in Combination

As described throughout this article, the framework of DSF offers numerous advantages for the examination of measurement invariance in polytomous items. But how should DSF be used in relation to traditional measures of DIF? Does the study of DSF nullify the need for traditional DIF indices? Or should DSF be used in a supporting role for the traditional analysis of DIF in polytomous items? We argue for the latter and recommend using indices of DIF and DSF in combination to garner the most sensitive and comprehensive information possible concerning the item. The test of DIF addresses the issue of whether there is a violation of invariance somewhere among the J steps, and the evaluation of DSF can provide rich information concerning the location and causes of the DIF effect.

As described in the previous section, global and net tests of DIF are relatively more sensitive to different forms of invariance; global tests of DIF are more powerful than net tests when the DSF effects are divergent or vary dramatically in magnitude across the steps, and net tests of DIF are more powerful than global tests when the DSF effects are relatively constant across all steps. As a result, we recommend using one global test of DIF (e.g., the SSL test or the generalized Mantel–Haenszel test) and one net test of DIF (e.g., the standardized mean difference, the Liu–Agresti cumulative common log-odds ratio, polytomous SIBTEST). If either of the global or net tests of DIF is significant, then an evaluation of the DSF effects should be pursued to explain the magnitude and cause of the DIF effect. This recommendation is consistent with that previously proposed by Penfield et al. (2009). Ultimately, we expect that using measures of DSF in combination with net and global tests of DIF can provide a more sensitive test of measurement invariance than using either approach in isolation.

An Example of a DSF Analysis

We will now illustrate the use of DSF and the associated interpretive issues presented earlier in this article using real educational achievement data. We conducted a DSF analysis on 30 polytomous items obtained from the 2001 Mathematics Assessment of the School Achievement Indicators Program (SAIP). The SAIP is a cycle of assessments in mathematics, science, reading, and writing developed by the Council of Ministers of Education, Canada, and administered across Canada between 1993 and 2004. These items primarily assessed mathematical skills in algebra, measurement, geometry, and data management. Data from English and French versions of the SAIP administered in two large Canadian provinces were analyzed. Of the 7,519 participants, 4,652 (61.9%) were administered the English version of the SAIP, and 2,867 (38.1%) were administered the French version. For the purpose of this study, students who took the French

version of the test were designated as the reference group, and students who took the English version were designated as the focal group.

An evaluation of DSF for each of the 30 polytomous items was conducted using the odds ratio approach (as described earlier in this article). The step-level common log-odds ratio ($\hat{\lambda}_j$), its estimated standard error, and the corresponding test statistic were computed using the DIFAS computer program (Penfield, 2005). Once the DSF estimates were obtained, the pattern of the DSF effects for each item was analyzed and interpreted. To demonstrate how the DSF analysis can be conducted in conjunction with the analysis of DIF, a global and a net test of DIF were also conducted for each item. A global test of DIF was conducted using the SSL test with a Bonferroni-adjusted Type I error rate (.05/*J*) for each step. A test of net DIF was conducted using the Liu–Agresti cumulative common log-odds ratio (*LA*), which was divided by its estimated standard error, resulting in a *z*-statistic that can be used as a net test of no DIF (Penfield & Algina, 2003). Although we implemented an odds ratio approach in the current analysis, logistic regression and IRT approaches could have been used and would be expected to yield results that are consistent with those presented here.

Six of the 30 items had a significant test of net or global DIF. Table 1 presents, for these six items, the DSF effect estimates ($\hat{\lambda}_j$), estimated standard errors, the associated *z*-statistics, the form of the DSF pattern, and the values of *LA* and its associated test statistic. Note that all six items led to a significant global test of DIF (the SSL test), and all but one item (item 13) led to a significant net test of DIF (the Liu–Agresti test). Also note that the DSF pattern was based purely on the magnitude of the DSF effects (i.e., not on the significance of the DSF effects).

Items 6, 7, and 15 were labeled as having pervasive DSF because all steps displayed a substantial DSF effect. The presence of pervasive DSF suggests that a biasing factor may exist at the item level. Item 6 displayed pervasive DSF that was constant, providing further evidence that the cause of the DIF effect resided in the item stem. Item 7 displayed a very large DSF effect for step 2 and a moderate DSF effect for step 1. This pattern suggests that the potentially biasing factor was more salient for the highest score category. Items 6 and 7 both required interpretation of the same graph of driving speed against time. The biasing factor for both items may be a difference in the availability and use of instructional technology (e.g., use of motion sensors and graphing calculators) among schools. Item 15, which required examinees to interpret the differences between two pie charts, displayed the largest DSF effects for the middle steps, suggesting that the biasing factor was most salient in the middle score categories, but the lack of a zero DSF effect in the highest score category suggests that the biasing factor is not restricted to the middle score categories. As a result, the DSF observed in item 15 appears to be attributable to an item-level effect.

Items 13, 16, and 30 displayed nonpervasive DSF, suggesting that the potentially biasing factors are localized to individual score categories. Item 13 displayed a large DSF effect only for the final step, indicating that the biasing factor resides in the highest score category, suggesting that

Table 1. Results of the DSF Analysis

Item	Step	$\hat{\lambda}_j$	*SE* ($\hat{\lambda}_j$)	*z* ($\hat{\lambda}_j$)	DSF Size	DSF Form	DIF Effect
Item 6	Step 1	.74	.14	5.33*	Large	Pervasive:	$LA = .76$
	Step 2	1.25	.50	2.49	Large	Constant	$z(LA) = 5.55$*
Item 7	Step 1	.52	.07	7.57*	Medium	Pervasive:	$LA = .54$
	Step 2	1.91	.51	3.76*	Large	Convergent	$z(LA) = 7.94$*
Item 13	Step 1	.01	.07	.12	Small	Nonpervasive:	$LA = .04$
	Step 2	−.03	.06	−.42	Small	Constant	$z(LA) = .83$
	Step 3	−.03	.06	−.49	Small		
	Step 4	.69	.13	5.55*	Large		
Item 15	Step 1	.77	.71	1.09	Large	Pervasive:	$LA = .41$
	Step 2	1.02	.59	1.73	Large	Convergent	$z(LA) = 8.12$*
	Step 3	.90	.43	2.11	Large		
	Step 4	.40	.05	7.80*	Small		
Item 16	Step 1	−1.12	.06	−20.42*	Large	Nonpervasive:	$LA = -.59$
	Step 2	−.26	.06	−4.08*	Small	Divergent	$z(LA) = -10.89$*
	Step 3	−.30	.07	−4.65*	Small		
	Step 4	1.32	.28	4.70*	Large		
Item 30	Step 1	1.31	.23	5.58*	Large	Nonpervasive:	$LA = .41$
	Step 2	1.17	.17	6.91*	Large	Constant	$z(LA) = 7.32$*
	Step 3	.85	.14	6.02*	Large		
	Step 4	.20	.05	3.75*	Small		

Note. The symbol $\hat{\lambda}_j$ corresponds to the cumulative step-level log-odds ratio estimator, $SE(\hat{\lambda}_j)$ corresponds to the estimated standard error of $\hat{\lambda}_j$, $z(\hat{\lambda}_j)$ corresponds to $\hat{\lambda}_j / SE(\hat{\lambda}_j)$, *LA* corresponds to the Liu–Agresti cumulative common log-odds ratio, and *z* corresponds to the *LA* divided by its estimated standard error. The asterisk (*) next to $z(\hat{\lambda}_j)$ indicates the values of $z(\hat{\lambda}_j)$ that exceeded the critical value associated with a standard normal distribution using the step-level Type I error rate of .05/*J* (i.e., a Bonferroni adjustment associated with a familywise Type I error rate of .05). The asterisk (*) next to $z(LA)$ indicates the value of $z(LA)$ that exceeded the critical value associated with a normal distribution using a Type I error rate of .05 (i.e., 1.96).

the biasing factor may well be in the scoring criteria for the highest score level. This item is categorized as having nonpervasive DSF that is constant (items exhibiting DSF in only one step are always labeled as having a nonpervasive constant pattern). It is also of interest to note that the net test of DIF (and the value of *LA*) was insensitive to the large DSF effect present in a single step. This provides a valuable example of the importance of considering DSF in combination with DIF: a traditional investigation of DIF using a net DIF index would have missed this lack of measurement invariance. Item 16, which required examinees to continue a linear pattern, displayed an instance of divergent DSF, such that the first step displayed a relative advantage for the English group, and the fourth step displayed a relative advantage for the French group. The DSF pattern observed in this item provides evidence of two distinct step-level effects favoring opposite groups at play in a single item. As a result, investigation of the potential biasing factor against the English group should focus on the highest score category (in line with the DSF effect of the fourth step), and investigation of the potential biasing factor against the French group should focus on the second highest score category (in line with the DSF effect of the first step). The French version of item 16 included a line reminding the examinees to show their work; the English version omitted this line. We would recommend beginning a search for the biasing factor by examining how whether students showed their work was considered in the scoring criteria for the top two score levels for this item. Item 30 displays a relatively constant DSF effect across the first three steps, indicating that a potentially biasing factor is localized to the second, third, and fourth score levels (i.e., not to the lowest and the highest score levels). Item 30 is the last question on the test, and the biasing factor is likely related to the differential speededness of the test in the two versions.

Concluding Remarks

Identifying and removing sources of bias from tests and test items is critically important to ensure that the tests are fair for all examinees. Although most test developers have well-established and rigorous procedures for identifying possibly biased dichotomously scored items, the available procedures for polytomously scored items have been less informative. Traditional approaches to detecting DIF in polytomous items may miss subgroup differences in difficulty that are specific to one or two score levels within a polytomous item and provide little information to help identify whether the source of the bias is in the item stem or in a particular aspect of the scoring criteria.

The possibility of failing to detect potential bias in polytomous items is particularly troubling when we consider that a polytomous item's contribution to an examinee's score is usually equivalent to several multiple-choice items. In effect, bias in a single score level within a polytomous item may have as large an impact as bias in a multiple-choice item, yet it may go undetected by traditional approaches.

Our goal in writing this article was to describe and demonstrate the advantages of DSF as a supplement to traditional DIF analyses of polytomous items. Already, one specialized DSF computer program (DIFAS; Penfield, 2005) exists, and this article provides the information test developers and researchers would need to perform DSF analyses using other computer programs, making DSF analyses feasible, even for large testing programs.

Appendix

Self-Test

1. Describe the difference between DIF and DSF.
2. There are two specific benefits of using the DSF framework in conjunction with the traditional DIF framework. Describe these two benefits.
3. A polytomously scored item has four score levels. How many DSF effects are associated with this item?
   (a) 2
   (b) 3
   (c) 4
   (d) 5
4. A DSF analysis indicates the following trend in DSF effects across the steps of an item: $\hat{\lambda}_1 = -.6,\ \hat{\lambda}_2 = 0,\ \hat{\lambda}_3 = .6$. What form of DSF best represents this pattern of DSF effects?
   (a) Nonpervasive divergent
   (b) Pervasive convergent
   (c) Nonpervasive convergent
   (d) Pervasive divergent
5. DSF analysis indicates the following trend in DSF effects across the steps of an item: $\hat{\lambda}_1 = 0,\ \hat{\lambda}_2 = .4,\ \hat{\lambda}_3 = .8$. What form of DSF best represents this pattern of DSF effects?
   (a) Nonpervasive divergent
   (b) Pervasive convergent
   (c) Nonpervasive convergent
   (d) Pervasive divergent
6. If the DSF effect for each step of a polytomous item is zero, then
   (a) DIF cannot exist for the item.
   (b) DIF must exist for the item.
   (c) It is possible, but not necessary, for DIF to exist for the item.
7. Suppose that a DSF analysis is conducted using the log-odds ratio approach. A colleague recommends that you also run a DSF analysis using the logistic regression approach. Would you expect to obtain similar results?
8. An IRT analysis is conducted and it obtains the following parameter estimates for the reference and focal groups: $\hat{a}_R = \hat{a}_F = 1.5,\ \hat{b}_{1R} = -1.5,\ \hat{b}_{1F} = -1.2,\ \hat{b}_{2R} = 1.0,\ \hat{b}_{2F} = 1.6$. Based on this information, what value best approximates the log-odds ratio DSF effect estimator for the second step ($\hat{\lambda}_2$)?
   (a) $\hat{\lambda}_2 = 0$
   (b) $\hat{\lambda}_2 = .3$
   (c) $\hat{\lambda}_2 = .6$
   (d) $\hat{\lambda}_2 = .9$
9. A DSF analysis indicates the presence of a constant pervasive DSF pattern. This pattern of DSF effects is most consistent with which of the following causes?
   (a) A problem associated with only one score level.
   (b) A problem associated with several, but not all, score levels.
   (c) A problem associated with the stem or prompt of the item.
10. Consider a polytomously scored item having score levels 0, 1, 2, 3, and 4. A DSF analysis of this item reveals the following DSF pattern: $\Delta(b_1) = 0$, $\Delta(b_2) = 0$, $\Delta(b_3) =$

.6, $\Delta(b_4) = 0$. Which of the following conclusions is most consistent with the obtained DSF effect estimates?

(a) The DSF is likely attributable to a property of score level 2.
(b) The DSF is likely attributable to a property of score level 3.
(c) The DSF is likely attributable to a property of score level 4.
(d) The DSF is attributable to the properties of score levels 1, 2, and 3.

Answers to Self-Test

1. DSF concerns whether between-group differences in each step function exist. Because the step functions determine the characteristic curve of each score level, DSF ultimately can be linked to which score levels are implicated in a DIF effect. In contrast, DIF concerns whether between-group differences in the score-level characteristic curves exist, but does not provide information about which score levels manifest the DIF effect.
2. Two benefits of DSF are: (a) DSF provides useful information for identifying the portion of the item responsible for a DIF effect, which can help test developers target potentially biasing factors; and (b) evaluating DIF using a DSF framework can be more powerful than traditional omnibus approaches for evaluating DIF if the DSF effects vary in sign and/or magnitude across the steps.
3. The correct answer is (b). The number of steps is equal to the number of score levels minus 1.
4. The correct answer is (a) because the DSF effects differ in sign and magnitude across the steps.
5. The correct answer is (c) because the DSF effects differ in magnitude, but not sign, across the steps.
6. The correct answer is (a).
7. Yes, the results are expected to be very similar. The DSF effect modeled using logistic regression is equivalent to the conditional log-odds ratio.
8. The correct answer is (d) because of the relationship $\lambda_j \approx a\Delta(b_j)$.
9. The correct answer is (c).
10. The correct answer is (b).

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Alvarez, K., & Penfield, R. D. (2007). *Using differential step functioning (DSF) to refine the analysis of DIF in polytomous items: An illustration*. Poster presented at the annual conference of the Institute of Education Sciences, Washington, DC.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*, 277–300.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, *37*, 307–327.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*, 333–353.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*, 335–350.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics, volume 26: Psychometrics* (pp. 979–1030). New York: Elsevier.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive assessment: Issues in constructed response, performance testing, and portfolio assessment* (pp. 135–166). Hillsdale, NJ: Erlbaum.

Douglas, J., Roussos, L., & Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement*, *33*, 465–484.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*, 315–332.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*, 164–187.

Hauck, W. W. (1979). The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics*, *35*, 817–819.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329–349.

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*, 345–355.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3–16.

Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, *52*, 1223–1234.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149, 174.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*, 284–291.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: Praeger.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Psychological Measurement*, *11*, 353–369.

Penfield, R. D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta Journal of Educational Research*, *49*, 231–243.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement*, *29*, 150–151.

Penfield, R. D. (2006). *A nonparametric method for assessing differential step functioning in polytomous items*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, *44*, 187–210.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, *40*, 353–370.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics, volume 26: Psychometrics* (pp. 125–167). New York: Elsevier.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), 5–15.

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*, 61–78.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*, 23–37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis: Detecting DIF item and testing DIF hypotheses. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107–115). Thousand Oaks, CA: Sage.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34* (4, Part 2, No. 17).

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement, 37* (1, Part 2, No. 18).

Samejima, F. (1997). Graded response model. In W. V. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement*, *24*, 97–118.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Erlbaum.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician*, *40*, 106–108.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53–75.

Thissen, D. (2001). IRTLRDIF (version 2) [Computer software]. Retrieved January 17, 2006, from http://www.unc.edu/~dthissen/dl.html.

Thissen, D., Chen, W.-H., & Bock, D. (2003). *MULTILOG 7* [Computer software]. Lincolnwood, IL: Scientific Software International.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Wang, W.-C., & Su, Y.-H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*, 450–480.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). *BILOG-MG 3* [Computer software]. Lincolnwood, IL: Scientific Software International.

Zumbo, B. D., & Thomas, D. R. (1996, October). *A measure of DIF effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.